HEART DISEASE DIAGNOSIS SYSTEM BY APPLYING CASE-BASED REASONING (CBR)

SAW SIEW CHING

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science (Software Engineering)

FACULTY OF COMPUTER SYSTEMS & SOFTWARE ENGINEERING
UNIVERSITI MALAYSIA PAHANG
KUANTAN, PAHANG

JUNE 2012

ABSTRACT

This is project overview of Heart Disease Diagnosis System (HDDS) by applying Case-Based Reasoning (CBR) technique. Diagnosis of disease is a vital and intricate job in medicine. There are a lot of Artificial Intelligence (AI) techniques to determine diagnosis. After comparisons were made between those techniques, CBR is chosen to diagnose heart disease. System presented in Java to make it accessible for users like health professionals and doctors. The result of this designed system to diagnose the heart disease effectively. HDDS developed using Eclipse as main compiler, SQL as database development tool, and Java language as the programming language. HDDS used Rapid Application Development (RAD) methodology in order to keep system more systematic.

ABSTRAK

Ini adalah gambaran keseluruhan projek Sistem Diagnosis Penyakit Jantung (HDDS) dengan menggunakan teknik Penaakulan Berasaskan Kes (CBR). Diagnosis penyakit adalah tugas yang penting dan rumit dalam bidang perubatan. Terdapat banyak teknik AI untuk menentukan diagnosis. Selepas perbandingan dibuat di antara kedua-dua teknik, CBR dipilih untuk mendiagnosis penyakit jantung. Sistem akan dibentangkan di dalam bentuk Java supaya pengguna seperti pakar kesihatan dan doktor mudah dan senang menggunakannya. Hasil daripada sistem ini akan dapat mendiagnosis penyakit jantung dengan berkesan. HDDS akan dikembangkan menggunakan Eclipse sebagai pengkompil utama, SQL sebagai alat pembangunan pangkalan data, dan bahasa Java sebagai bahasa pengaturcaraan. HDDS menggunakan metodologi Pembangunan Aplikasi Pantas (RAD) untuk memastikan sistem yang lebih sistematik.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
CANFIS	Coactive Neuro-Fuzzy Inference System
CASE	Computer-Assisted Software Engineering
CBR	Case-Based Reasoning
CHD	Coronary Heart Disease
GA	Genetic Algorithm
HDDS	Heart Disease Diagnosis System
HDL	High Density Lipoprotein
JRE	Java Runtime Environment
mg/dl	milligrams per deciliter
mmHg	millimeter of mercury
NN	Neural Network
PSM	Final Year Project
RAD	Rapid Application Development
SCBR	Statistical Case-Based Reasoning
SQL	Structured Query Language
SVM	Support Vector Machine
UMP	Universiti Malaysia Pahang

CHAPTER 1

INTRODUCTION

This chapter briefly describes the Case-Based Reasoning System for Support Diagnosis of Heart Disease. In this chapter, there are five sections which include the introduction of the project and problem statement related to this project. Third section is the objectives to be achieved and fourth section is the scopes. Last section is the thesis organization.

1.1. Introduction

The term 'Heart Disease' encompasses the diverse diseases that affect heart. The number of people suffering from heart disease is on the rise (Health Topics, 2010). The report from World Health Organization shows us a large number of people that die every year due to the heart disease all over the world. Heart disease is also stated as one of the greatest killer in Africa, America and also Asia.

Correct diagnosis of heart disease at an early stage is a demanding task due to the complex interdependence on various factors. Another major challenge faced by hospital is the provision of quality services at affordable cost. These are the motivations to develop medical diagnosis prediction system which can predict heart disease by processing the previous known cases.

Working on heart disease patients databases is one kind of a real-life application. The databases include several factors and attributes. Therefore, CBR was proposed for supporting diagnosis of heart disease in this study. CBR, this method in Artificial Intelligent is applied in this system by a real medical data set taken from UCI machine learning. David W. Aha, the data set donor has mentioned that there were some papers that focused on the data by different techniques (Aha, 1988). There are some papers using different technique to predict or deduce heart disease, such as Fuzzy Expert System, Multilayer Perception and also using Coactive Neuro-Fuzzy Inference System (CANFIS) and Genetic Algorithm.

CBR has been used in various problem-solving areas such as financial forecasting, credit analysis and medical diagnosis. In addition, CBR is chosen because CBR is appropriate in medicine for some important reasons: cognitive adequateness, explicit experience, duality of objective and subjective knowledge, and system integration. CBR is unlike the traditional rule-based approach in which expert knowledge must be represented in “if-then” rules, CBR manages attributes to be grouped and stored.

1.2. Problem Statement

Heart disease is the second leading killer disease in Malaysia. It is also a major cause of disability (Health Topics, 2010). Generally, doctors and health professionals use their knowledge and experience to make decision for the diagnosis of heart disease for patients. Usually, most of the medical data collected from patients are just saved in files or kept in folders. Generally, those huge amounts of messy medical records have not meaning for users. Using CBR, a technique which solves a new problem by remembering a previous case and by reusing information and knowledge of that case, CBR turn those data into useful information that can help to make decision support system for the diagnosis of heart disease.

This system can be used to assist doctor and support education for the undergraduate and postgraduate young physicians as a tool to improve the quality of care for the patients. This system can be used as a reference for those student and new doctor. Presently, doctors have difficulties in determining heart disease in a new patient who does not have existing medical record. Therefore, those data can be used to diagnose heart disease for new patients who do not have existing medical records.

This system is designed to assist doctor and health professionals in determining the diagnosis of patient data. Therefore, this system could help doctors and health professionals to determine the diagnosis and analysis of the patient health status.

1.3. Objective

Objectives of this study are:

i. To develop an intelligent clinical decision support system for the diagnosis of heart disease.

ii. To use CBR algorithm to predict heart disease in patient.

1.4. Scope

The way of doctors and health professionals diagnose heart disease depends on their experience and knowledge. That way is similar to decision support system using CBR approach. CBR algorithm has 4 phases which are retrieve, reuse, revise and retain. But this system only uses retrieve and reuse technique in whole cycle of CBR algorithm. In addition, this project retrieves the dataset from UCI Machine Learning, and reuses its database to support the system. Secondly, this system only can support prediction for diagnosis of heart disease using same attributes in the system database. Limiting the amount and areas of attributes can help this system in more constant way. To help more users' access to the system, accessibility is needed and applied in this project.

1.5. Thesis Organization

This system presents CBR for supporting diagnosis of heart disease and organizes those five chapters as follows:

Chapter 1 : Introduction of this project.

This chapter involves the introduction on the project. It was included elaboration of problem statement, scopes and objective of this project.

Chapter 2 : Literature Review

It presents the literature review among CBR and techniques supporting diagnosis of heart disease and also deal with mobile operating system and advantages of choosing CBR.

Chapter 3 : Methodology

The section where methodology utilized in simulating the research project is illustrated in.

Chapter 4 : Expected Result and Discussion

This chapter presented the actual implementation of the research project.

Chapter 5 : Conclusion

It involves in reviewing the findings & results and discusses the outcome of the project.

CHAPTER 2

LITERATURE REVIEW

2.0 Literature Review

Literature review of this project presented about the technique and equipment that are going to be used in this project.

2.1 Heart Disease

Heart disease is a class of diseases that involves the heart or blood vessels. Heart disease is the second leading cause of death in Malaysia for men and women. There are many different forms of heart disease. The most common cause of heart disease is narrowing or blockage of the coronary arteries which are the blood vessels that supply blood to the heart itself. This is called coronary artery disease and it happens slowly over time. It is the major reason people have heart attack.

Other types of heart problems may happen to the valves in the heart, or the heart may not pump well and causes heart failure. There are some people who are born with heart disease.

Many things increase the risk for heart disease, and mostly people want to reduce those risk factors. In this case, the factors are:

- Having diabetes which is a strong risk for heart disease.

- Substance abuse such as cocaine
- Being overweight
- Not getting enough exercise and feel depressed or having excess stress
- Smoking
- High blood pressure increases the risks of heart disease and heart failure
- Excess cholesterol in blood build up inside the walls of heart's arteries (blood vessels)

People can help to reduce the risk of heart disease by taking steps to control factors mentioned, for example by controlling the blood pressure, lowering the cholesterol level, refraining from smoking and having enough exercise.

2.1.1 Table of Heart Disease Database

The heart disease data set was supplied by R. Detrano, PhD of the V.A. Medical Centre. This data set can be found in Cleveland Clinic Foundation (Park *et al*, 2006) and UCI Machine Learning (Aha, 1988). The database contains 270 set of cases and 76 attributes, and only 13 attributes are meaningful in this study. The database consists of two classes of result where 150 cases are absent of heart disease and 120 cases show the presence of heart disease. This study only use the 13 attributes to diagnose heart disease and compare result with the result given.

Table 2.1: Heart Disease Attributes and Description

	Attributes	**Description**	**Values**
1	Age	Age of patients in year	Integer
2	Gender	Gender of patients	Male/female
3	CP	Chest pain type Angina: typical angina Abnang: Atypical angina Notang: Non-angina pain Asympt: Asymptomatic	Four types
4	Trestbps	Resting blood pressure in mmHg on admission to the hospital	Integer
5	Chol	Serum Cholesterol in mg/dl	Integer
6	FBS	Fasting sugar pressure>120 mg/dl True = 1; False = 0	0, 1
7	RestEcg	Resting electrocardiographic result 0: Normal 1: having ST-T wave abnormality 2: showing probability or definite left ventricular hypertrophy	0 – 2
8	Thalach	Maximum heart rate achieved	Integer
9	Exang	Exercise included angina True = 1; False = 0	0, 1
10	OldPeak	ST depression Included by exercise relative to rest	Float
11	Slope	Slope of peak exercise ST segment 1: up sloping 2: flat 3: down sloping	1 – 3
12	CA	Number of major vessels colored by fluoroscopy (0-3)	0 - 3
13	Thal	Defect type 3: Normal 6: Fixed defect 7: reversible defect	3, 6, 7
14	Result	Heart Disease buff: Healthy sick: Sick	buff, sick
Patient ID		Patient's identification number	Integer

2.2 Exists System

In the medicine field, several diagnostic tools for heart disease were developed in different AI techniques.

2.2.1 Statistical Case-Based Reasoning Expert System: Application to Medical Diagnosis (Park *et al*, 2006)

In the research of those authors, Statistical Case-Based Reasoning (SCBR), this new knowledge extraction technique suggested in SCBR roles to adapt optimal number of neighbours dynamically by considering the distribution of the distances between cases. It was improved from CBR and SCBR where selection of the case is based on the degree of similarity between the potential of neighbours in each other case that was presented by the authors in this paper. Step one of SCBR was to scale data then learn the distribution of distances of the database. Then, the authors found out the optimal cut-off probability threshold by calculation and the last step is to perform CBR by selected neighbours and calculate the result.

The database used in this system is the same as the database used in this study. This system implemented 13 attributes in SCBR to diagnose heart disease and compare the result with the final attribute (Aha, 1988).

To understand the distribution of distances, the distance of pair wise between cases were calculated and transformed into normal distribution. It used statistic knowledge to understand the confidence and get information from the calculation. The optimal cut-off probability calculated needed to have different optimal distance threshold between cases. The last step was the implementation of CBR into the calculation. The accuracy of SCBR and CBR were listed in the system and it would be a good review for researchers to compare the difference. This was because of this existing system did not weight the attributes of heart

disease thus the result and accuracy of the system would be different with this study.

Table 2.2: Summary Result of Sensitivity and Specificity between CBR and SCBR

Dataset	**Measurement**	**CBR**	**SCBR**
Heart Disease	Sensitivity	77.50%	72.50%
	Specificity	85.33%	92.67%

2.2.2 Fuzzy Expert System for Determination of Coronary Heart Disease Risk (Allahverdi *et al*, 2007)

This existing system has been supported by Selcuk University's Scientific Research Unit. In this paper, researcher found out that the cholesterol level has been identified as one of the main risk factor for the Coronary Heart Disease (CHD). The risk assessment to determine the 10 years risk is dependent on the Framingham Risk Scoring. From the appendix, the 5 attributes have been pointed out: age, total cholesterol, High Density Lipoprotein (HDL) cholesterol, blood pressure, treatment for hypertension and cigarette smoking. According to the Framingham Risk Scoring, authors assumed that the age, cholesterol level, HDL cholesterol and blood pressure level affect the CHD risk.

The first step of the system was to calculate the number of points for each risk factor. After that, the blood pressure was counted into the system at the time of assessment to determine whether the person was needed to be on anti-hypertensive therapy. The cholesterol level remained as the primary target of therapy. Besides, other attributes calculated as general points. The total point scored for those risk factors compared to the 10 years risk as indicated in the Appendix I. The calculation of this system showed that the next 10 years of CHD risk by the factors above.

The system categorised risk in 3 classes that have CHD and CHD risk equivalents (10 years risk > 20%), multiple risk factor (10 years risk 10-20% and 10 years risk < 10%), and 0-1 risk factor. Once the input values reached the target of the classes, output of determination displayed to the user. System recommend three outputs in normal living, which is the 0-1 risk factor in the calculation, diet or grog treatment for the CHD and CHD risk equivalents patients.

2.2.3 Heart Disease Prediction System using Coactive Neural-Fuzzy Inference System and Genetic Algorithm (Parthiban & Subramanian, 2007)

Coactive Neural-Fuzzy Inference System (CANFIS) is a model that integrates adaptable fuzzy inputs with a modular Neural Network (NN) to rapidly and accurately approximate complex function. CANFIS solves problem more efficiently than traditional NN when the underlying function to model is highly variable or locally extreme. In addition, CANFIS is useful in selecting the most relevant features of the data which can produce a smaller and less complicated network. The advantage of the combination of Genetic Algorithm (GA) with other AI techniques is to help the system operates with the goal of finding the best solution to a problem by searching until the specified criterion is met.

The database of this system came from UCI Machine Learning (Aha, 1988). This system implemented 13 attributes by using CANFIS with GA to predict heart disease.

Crossover in GA needs chromosomes randomly paired and segments of paired chromosomes between two randomly determined breakpoints were swapped. It could be inverted and become incorporated into the recipient. In this system, GA used the serial method of binary type, roulette-wheel in the selection operator, and boundary in the mutation operator. All the chromosomes were automatically set in the system and so that the system consisted of the number of input and membership functions, learning rate and momentum. The reason for the high accuracy was due to the same set of data that were used to train and test. The system only chose 153 sets of data randomly and tested the algorithm with very high accuracy which was 98.40%.

2.2.4 Decision Support System by Using Multilayer Perception (Godara & Nirmal, 2010)

This paper explains the supervised feed forward neural network and the usage of back propagation learning algorithm with momentum term and augmented learning rate. In this paper, researchers mentioned that data mining has been heavily used in medical field to include patient diagnosis records to identify the result of the patients. Multi-layer perception is a branch from neural network with a supervised learning algorithm and back propagation. Researchers also provided the result for comparison of the accuracy for multi-layer perception and multi-layer perception with Dagging approach.

Data source for this system is the same as the system in Section 2.2.1 and Section 2.2.3, which was taken from UCI Machine Learning.

Table 2.3: Result of Accuracy, Sensitivity and Specificity between Multilayer Perception and Multilayer Perception with Dagging Approach

	Accuracy	**Sensitivity**	**Specificity**
Multi-Layer Perception	81.85%	84.24%	78.99%
Multi-Layer Perception with Dagging Approach	84.58%	86.67%	81.16%